

**A COMPARISON BETWEEN RIGHT AND WRONG
ANSWERS ON A MULTIPLE CHOICE TEST**

J. C. POWELL

**University of Windsor
Windsor, Ontario**

ALVIN G. ISBISTER

**Red Deer College
Red Deer, Alberta**

A COMPARISON BETWEEN RIGHT AND WRONG ANSWERS ON A MULTIPLE CHOICE TEST

J. C. POWELL

University of Windsor
Windsor, Ontario

ALVIN G. ISBISTER

Red Deer College
Red Deer, Alberta

A test was developed so that right and wrong answer subtests could be compared. A clear pattern emerged which supported the a priori construct validity assumptions. The results were sufficiently conclusive to apparently *refute* the assumption often made or implied that "wrong answers contain no achievement information not included in the right answers or subtests thereof."

In an earlier paper Powell (1968) reported the results of the a posteriori analysis of wrong answers on Gorham's (1956) Proverbs Test. The results of that study suggested that the assumption that wrong answers are randomly distributed is untenable. In this present paper, factor analytic techniques are used to study some of the characteristics of the a priori classification of both right and wrong answers to investigate as to how right and wrong answers might be related.

Design of the Test

An experimental test was designed especially for this study. In order to provide the students with a common background a set of five short reading selections was supplied in the test. These selections were chosen on the basis that it was unlikely for freshman education students to have encountered them before taking the test. The multiple choice questions were based on the information

content of single selections or selections in various combinations. An attempt was made to have the items require no factual information which was not included in the selections except for a general understanding of the vocabulary of the selections and items.

Alternatives were classified on three bases: (1) Bloom's *Taxonomy* (1956) for the correct answers; (2) the logical structure of the wrong answers as they related to the correct ones, and, (3) logical fallacies based on Chase's *Guides to Straight Thinking* (1956). These latter classifications were arbitrary. They did not provide categories of classification for all the wrong alternatives. Both the right and the wrong classifications were to establish an a priori multiple-keying system which was used to determine a total correct score, and subtest scores, for both "right" and "wrong" answers.

Table 1 illustrates the manner in which the items for this test were developed.

TABLE 1

FIRST READING SELECTION

Source: Dexter, Lewis Anthony: *The Tyranny of Schooling*; N.Y., Basic Books, 1964, p. 1.

Most people in our society at one time or another suffer humiliation, shame, or at least severe apprehension because of one great fear: they are afraid that other people may think that they are stupid. This fear of being regarded as stupid frequently underlies inferiority complexes, self-contempt, self-depreciation, and despair.

Our society teaches contempt for stupidity and fear of being regarded as stupid through one central institution and its auxiliaries. This institution is compulsory schooling. It is aided by such auxiliary practices as compulsory written examinations for admission to many jobs, intelligence testing, and the like.

SECOND READING SELECTION

Source: Marris, Peter: *The Experience of Higher Education*; London, Routledge Kegan Paul, 1964, p. 175.

In this sense, it does not matter what subject a student studies, since each is leading toward a generalized intellectual awareness. But the starting point is still important since a student has the greatest incentive to understand whatever relates most immediately to his interests. Nor are the concepts derived from any one field of study equally relevant to any others: The ramification of insights remains biased by its roots. The intellectual content has to both guide and be guided by the purpose for which a student seeks understanding. Otherwise it is meaningless.

If, then, higher education aims to teach students how to abstract, from a particular context, principles by which they can organize the perception of their universe of thought, it requires that these students have a use for such free-ranging understanding. When they enter higher education, their aims are confused, and they may not see, or wish to see, the value of a generalized intellectual skill. Their approach to learning has been conditioned by extraneous motives: They worked to win approval or avoid blame, to pass an examination, as much or more than for the sake of understanding. They are not used to asking themselves what they want to understand, or why, but derive enough interest to master the skills required of them

TABLE 1 (Continued)

from a desire to satisfy the authority who sets the task. So, I think, the function of higher education is as much to develop the autonomy of their desire to understand, as to satisfy it.

THIRD READING SELECTION

Source: Kagan, J.; Moss, H. A.: *Birth to Maturity*, N. Y., Wiley, 1962, p. 85.

Aggression is a second behavior system that begins its growth during the first five years. Traditionally a response was labelled aggressive if the goal of the behavior was assumed to be psychological or physical injury to a person or person surrogate. We have adhered to this definition. As with dependency, the display of aggressive acts is a regular concomitant of development. The slapping or pushing of an age mate, the destruction of a sibling's new fort and the stinging verbal attack are regularly observed in the behavior of many children.

Aggression, like dependency, is subject to socialization pressures, for the child does not have complete license to unleash his anger when he chooses. In addition, as with dependency, the occurrence of overt aggression is a function of both the threshold for motive arousal and the intensity of anxiety associated with direct expression of this behavior.

In contrast to dependency, however, the potential for conflict over aggression is greater for females than for males. The pattern of social rewards and traditional sex-role standards act in concert to discourage the direct expression of aggression in girls and women. It might be anticipated, therefore, that aspects of aggression would be more stable for males than for females. This is precisely what occurred, for overt aggression to mother and frequent tantrums during childhood predicted adult aggressivity for men but not for women.

ITEM 1

Bloom's Class: Synthesis 5.30

Assuming that the school approves of contempt for stupidity, then:

- (a) aggression would increase and autonomous thinking increase.
- (b) aggression would decrease and autonomous thinking decrease.
- (c) aggression would increase and autonomous thinking decrease.
- (d) aggression would decrease and autonomous thinking increase.

ITEM 2

Bloom's Class: Evaluation 6.10

Which of the following statements is correct:

- (a) aggression is definitely undesirable from the standpoint of education.
- (b) aggression is potentially useful for educational purposes.
- (c) aggression can be eradicated from an individual's behavior.
- (d) aggression must be relieved through a particular mode of behavior.

From Table 1 we can examine the information background supplied for each of the two sample items selected as representative of the entire thirty items on the test. Comments on each item follows.

Item 1. This item is classified as Synthesis (5.30) because more than one of the selections are involved in the answer.

Assuming the information from the above articles to be true, the correct answer must be alternative (c). Alternative (a) is incorrect because autonomous thinking would not be expected to increase in an environment of external negative (aversive) reinforcement. Alternative (b) is incorrect due to the fact that aggression would be expected to increase in these conditions. Alternative (d) is at fault since it is the exact opposite of the correct answer.

If we represent an increase in aggression as A and an increase

in autonomous thinking as B, then the foils can be represented symbolically as:

- a. $A \cap B$
- b. $(\sim A) \cap (\sim B)$
- *c. $A \cap (\sim B)$
- d. $(\sim A) \cap B$

These four alternatives thus represent all the possible binary conjunctive combinations of two propositions and their inverses. The alternatives are therefore considered to be produced on the basis of logical structure.

Item 2. This item is classified as Evaluation (6.10) because of the value judgment required in the question. This item has the foils established on the basis of logical fallacies. These statements are essentially hypotheses which can be tested with reference to the Third Reading Selection. It should be noticed that in order to select foil a, the examinee will have to assume that aggression is invariably harmful. This assumption is clearly not made in the excerpt. For this reason, a person selecting this foil must begin with a Invalid Major Premise.

Foil c has two characteristics which distinguish it. It makes the same faulty assumption as foil a, and adds to this assumption an element of wishful thinking. Because it is worded emotionally ("eradicate") as well, it is classified as a Vox Populi Fallacy.

In foil d, the conclusion simply does not follow from the information given. Hence it is treated as a Non Sequitur Fallacy. Thus by either eliminating the "wrong" answers or by recognizing the role of social pressures in channeling aggression the student should be able to arrive at the correct answer (b).

For all 30 items the foils (wrong alternatives) were designed in either of these two modes.

Hypotheses Tested

There are several questions raised by the design of the test and its analysis which need clarification.

First: Are these arbitrary classifications mutually exclusive? Bloom's *Taxonomy* is presumed to be a subsumptive hierarchy. If it is, so far as this test is concerned, the relationships between the

right answer subtests should be oblique, making second order factors possible. A promax rotation was used to test this possibility.

Second: All alternatives which received less than five per cent selection ratio and those subtests containing less than five members were dropped from the final analysis of the responses in order to assure meaningful results among the factors. Does the fact that the subtests contain linear dependencies with respect to the total correct score influence the interpretability of the factors, or is the dropping of four alternatives (out of 30) from the right answer subtests, and 48 alternatives (out of 90) from the wrong answer part of the analysis sufficient to remove these dependencies? The rank of the correlation matrix was found to answer this question.

Third: Since more than one alternative from each item was used in the analysis, do the experimental dependencies which this fact introduces influence the results of the analysis? The item-by-item overlap between subtests was examined to answer this question.

Fourth: Does the information from the analysis contribute anything useful or interesting to the field of test construction?

Procedure

This test was administered to 307 first year university students in an introductory course in educational psychology.

The results from all subjects were scored for total correct and both "right" and "wrong" answer subtests. The right answer scores included the whole test score (out of 30) and three subtests based on Bloom's *Taxonomy: Comprehension* (8); *Analysis* (11); and *Synthesis* (7). Wrong answer subtests were five in number: *Vox Populi* (13); *Non Sequitur* (11); *Over Generalization/Over Simplification* (8); *Irrelevancy* (5); and *Invalid Major Premise* (5). The number of alternatives in each subtest is indicated in parentheses in each case.

These scores were treated as independent and subjected to a principal components factor analysis with one (1.00) in the diagonal followed by varimax and promax rotations.

Results

Table 2 shows the varimax rotation, and the transformation matrix. The promax rotation did not improve the structure when the fourth power was used, hence it is assumed that the classifications

TABLE 2
Combined Right and Wrong Answer Factor Pattern (Varimax Rotation)

Variable	Communality	I	II	Factors III	IV	V
Whole Test	.94	.89 ^a	—	—	—	—
<i>Right Answer Subtests</i>						
Comprehension	.65	—	.68	—	—	—
Analysis	.69	.79	—	—	—	—
Synthesis	.74	—	—	.75	—	—
<i>Wrong Answer Subtests</i>						
Vox Populi	.82	—	—	—	.60	.50
Non Sequitur	.76	—	—	—	-.81	—
Over Gen/Over Simp	.81	—	—	—	—	-.85
Irrelevancies	.55	—	.70	—	—	—
Invalid Major Prem	.58	—	—	-.71	—	—
<i>Transformation Matrix</i>						
	0.895	0.279	0.283	0.141	0.144	
	0.192	-0.405	-0.637	0.547	0.307	
	-0.245	0.854	-0.308	0.212	0.265	
	0.161	-0.044	-0.340	-0.771	0.511	
	0.276	0.161	-0.551	-0.202	-0.744	

^a Only factor loadings >|.38| are shown.

are mutually exclusive, and that no second-order factors are determinable.

The use of Gaussian elimination on the correlation matrix given in Table 3 gave a rank for this matrix of nine which is equal to the number of variables. For this reason, we may conclude that there are no linear dependencies in this data.

Finally, the alternatives composing each subtest are reported in Table 4. It should be noted that the only factor for which the pattern is strongly dependent on the overlapping of items in different

TABLE 3
Correlation Matrix

	W.T.	Comp.	An.	Syn.	V.P.	N.S.	Irr.	OS/OG	I.M.P.
Whole Test.	1.000								
Comprehension	.494	1.000							
Analysis	.620	.089	1.000						
Synthesis	.537	-.006	.039	1.000					
Vox Pop.	-.215	-.114	-.052	-.077	1.000				
Non Seq.	-.218	-.175	-.157	-.054	-.090	1.000			
Irr.	-.015	.089	-.075	.028	-.064	-.008	1.000		
OS/OG.	-.196	-.167	-.144	-.074	-.054	-.018	-.040	1.000	
Inv. M.P.	-.057	-.066	.081	-.090	.051	-.034	-.014	-.033	1.000

TABLE 4
Membership of Alternatives in Subtests

Subtest	Item																														Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Analysis			x		x				x									x	x	x	x	x	x	x		x					11
Comprehension	2	x		x		x	2					x			x		x														8
Synthesis											x			x											3		x	x	x	x	7
Vox Populi	x		x	4	5	x			4	4		4	4	4				4							x	4					13
Non Sequitur				4			x		4	4		4	4	4	x			5								5					11
O-G/O-S					5		x	x									5								4	5		x	x		8
Irrel.	2			x	x		2																		x						5
I.M.P.	x					x			x				x												3						5

x is item in subtest.

Numbers represent factors of overlapping items in related variables.

SAMPLE ITEM 1 is Number 11 above.

SAMPLE ITEM 2 is Number 13 above.

POWELL AND ISBISTER

subtests is Factor IV. The two variables strongly represented in this factor (Vox Populi and Non Sequitur) have eight items in common. The two variables in Factor II have two items in common and since these items would contribute to an opposite polarity of loadings, and the actual polarity is identical, some reason other than experimental dependencies must account for this pattern. A similar argument may be used for Factors III and V as well which have one and three common items respectively.

Discussions and Conclusions

The factor pattern shown has an unusually clear structure with large portions of the communalities accounted for by most of the factor loadings shown. The test appears to be mainly an analysis test in the a posteriori factor pattern, which confirms the a priori design discussed above.

The distribution of wrong answer subtests on the right answer subtests is very interesting. There is, for instance, a distinct and strong polarity between the Synthesis and Invalid Major Premise subtests.

This polarity would suggest that the reliability of the Synthesis subtest scores could be increased by *subtracting* each individual's Invalid Major Premise subtest scores from them. This procedure would effectively lengthen the Synthesis subtest from seven items to 11, thus increasing the reliability.

Another interesting polarity exists between Vox Populi on the one hand and the combination of Non Sequitur and Over Generalization/Over Simplification on the other. These foil responses may display this pattern in Factor IV because of the overlap already discussed. However, for this overlap to produce the observed results in Factor IV the selection of these mutually exclusive alternatives must be produced by an interlocking response pattern of a relatively well defined subgroup of the examinees. (This overlap is illustrated in sample item 2.) None of the other Factors display this experimental dependency. The fact that the Vox Populi alternatives distribute themselves on this particular pair may be, therefore, logically significant.

It is also interesting that Irrelevancies load positively on the same factor as Comprehension. Analysis of the distribution of selection of this class of foil shows that it is most frequently selected by the

middle group and the highest scoring group when total correct score is used as a scale. This result is consistent with previous findings. (Powell, 1968). The Comprehension subtest might therefore be lengthened by including Irrelevancy answers in it. Also, as indicated elsewhere, (Powell, 1968) a high Irrelevancy subtest score combined with a high total correct score may help to distinguish divergent thinkers from convergent thinkers, wherein the latter would have a low Irrelevancy score and a high total correct score. This observation might be used to overcome the criticism that multiple choice tests cannot be used to identify divergent thinkers.

It is fairly evident that a priori and a posteriori classification of foil selection yielded consistent results.

Unfortunately, of the two foil generating procedures used, only the logical fallacies procedure was studied in the results. Among these, this procedure produced foil subtests which were, in general, self-consistent to a degree mutually exclusive, and related in an apparently reasonable manner to the right answer subtests and to the findings of other research. It is therefore apparent that the logical fallacies procedure may well be an effective approach to foil generation when constructing higher mental process multiple choice tests.

A study of foils using the logical structure approach may yield comparable results if the categorization problems can be overcome. Certainly the selection of foils seems to be non-random in this case as well as in the other one cited (cf. Powell; 1968).

In conclusion, four implications seems to emerge from this study. First, logical fallacies appears to be a useful method for the development of foils (wrong alternatives) when constructing higher mental process tests.

Second, by using both right and wrong answer scores, it may be possible to improve the reliability of the right answer subtests, and possibly the construct validity of the entire test.

Third, it may be possible to obtain achievement information (such as a distinction between divergent and convergent thinkers) from multiple choice tests by using right and wrong answers in conjunction, that may not be available from total correct scores or right answer subtest scores alone.

Fourth, as indicated earlier, (Powell, 1968) found contradiction to the assumption that wrong answers are selected by the students

on a "random" as contrasted with a "systematic" basis. This present study carries this challenge to the traditional approach to wrong answers one step further. The usual assumption which is made is that, random or not, wrong answers are somehow "opposite to" right answers, and as such are linearly related. If this assumption were true, wrong answers would not contain any information not present in right answers. The findings of this study suggest that this approach to wrong answers may be in error. The following summary makes this evident:

1. No "opposites" to Analysis were found in this study (although the absence of logical structure foils may account for this).
2. Not all wrong answer subtests were "opposite" in factor polarity to right answer subtests. (Comprehension and Irrelevancies).
3. Using nearly twice as many wrong alternatives as right alternatives in this study did not produce evidence for linear dependencies (Gaussian elimination) or otherwise collapse the space (Promax).
4. There were wrong alternative subtests which were independent of any right answer subtests (Factors IV and V).
5. Further support to the possible relationship between total score, and Irrelevancy selection on the one hand and divergent thinking on the other was found.

As a test for the assumption that wrong answers contain no useful information which cannot be derived from right answers (either in total, or in subtest) this present study may well be definitive. This study appears to contain enough contrary evidence to refute this assumption. At least, these findings indicate that further study is needed into what may prove to be a new area in achievement test construction.

REFERENCES

- Bloom, B. S. *Taxonomy of educational objectives: Handbook I, cognitive domain*. N. Y.: David McKay, 1956.
- Chase, S. *Guides to straight thinking with 13 logical fallacies*. N. Y.: Harper, 1956.
- Dexter, L. A. *The tyranny of schooling*. N. Y.: Basic Books, 1964.
- Dinkmeyer, D. C. *Child development*. Englewood Cliffs, N. J.: Prentice-Hall, 1965.
- Gorham, D. R. *Proverbs test*. Missoula, Montana: Psychological Test Specialists, 1956.

- Kagan, P.** *The experience of higher education*. London: Routledge Kegan Paul, 1964.
- Powell, J. C.** The interpretation of wrong answers from a multiple-choice test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 403-412.
- Presscott, D. A.** *The child in the educative process*. N. Y.: McGraw-Hill, 1957.

